

Sparse Exponential Family Latent Variable Models

Shakir Mohamed¹, Katherine Heller and Zoubin Ghahramani

Department of Engineering, University of Cambridge

Recent efforts in high dimensional data modelling have highlighted the need for models that encode sparsity. In particular, latent factor models with sparsity have become relevant in situations where we believe that there are a number of underlying factors but that only a few factors are active and contribute to explaining any particular data point. One clear example where a sparse representation is applicable is in gene expression modelling. A gene’s expression is influenced by the presence of a number of transcription factor proteins, and there exists a wide array of such transcription factors that may affect the expression of any set of genes. Here, the underlying biology is sparse, since an individual gene’s activity may only be directly influenced by a small subset of the known transcription factors.

We consider sparsity in the framework of generalised latent variable models. These models are based on an exponential family likelihood and are generalised models in a manner analogous to GLMs for regression. We specify the model for observations $\{\mathbf{x}_n : n = 1, \dots, N\}$ as follows:

$$\mathbf{x}_n | \mathbf{v}_n, \Theta \sim \text{Expon} \left(\sum_k v_{nk} \boldsymbol{\theta}_k \right) \quad \boldsymbol{\theta}_k | \boldsymbol{\lambda}, \nu \sim \text{Conj}(\boldsymbol{\lambda}, \nu) \quad \mathbf{v}_n \sim \prod_{k=1}^K \mathcal{S}(v_{nk} | \varphi) \quad (1)$$

The conditional distribution of the observed data \mathbf{x}_n , is any exponential family distribution denoted $\text{Expon}(\cdot)$, where the natural parameters are a sum of the parameters $\boldsymbol{\theta}_k$, weighted by v_{nk} , the points in the latent subspace corresponding to data point \mathbf{x}_n . The parameters $\boldsymbol{\theta}_k$ are modelled using the corresponding conjugate distribution, denoted $\text{Conj}(\cdot)$ with hyperparameters $\boldsymbol{\lambda}$ and ν . A priori, the K latent variables for each observation, $\{v_{nk} : k = 1, \dots, K\}$ are assumed to be independent.

We study these models using sparsity-favouring priors $\mathcal{S}(v_{nk})$ for the latent variables. A sparsity favouring prior is any distribution with high excess kurtosis, indicating that it is highly peaked with heavy tails, or a distribution with a delta-mass at zero. We explore in depth various sparsity favouring priors including continuous priors such as the Laplace or Exponential distribution, and “spike and slab priors” with a delta mass at zero. Continuous sparsity favouring priors allow for sparse learning but place no mass on zero itself and thus samples are never exactly zero. Spike and slab priors have the desirable property that posterior samples contain zeroes, but have thus far been relatively unexplored in the unsupervised setting.

We consider Bayesian models using a sampling approach to inference with a Laplace, Exponential and spike and slab prior. We compare these models to a commonly used optimisation approach based on L_1 norm regularisation. Experimental results show that the “spike and slab” model has the best performance on data reconstruction as measured using the predictive probability on held-out data as well as the root mean squared error. Evaluations are shown on both synthetic data generated from the model as well as three real world data sets consisting of human judgements of animal features, robot planning and SPECT images.

In compressed sensing and related areas, the idealised, but intractable optimisation criterion uses an L_0 norm to penalise the number of non-zero parameters. The spike and slab model can also be seen as placing a penalty on the non-zero parameters, and thus can be seen to enforce sparsity in a manner similar to an L_0 norm minimisation. The spike and slab model approach has the property that it is able to introduce sparsity, while not enforcing shrinkage on parameters where no shrinkage is necessary, as is the case with models which use the Laplace prior for example. This allows for accurate data reconstruction, while learning the appropriate sparsity pattern supported by the data. The results are extremely encouraging and suggests a much wider scope for the use of spike and slab models in Bayesian unsupervised learning settings.

¹Presenting author (sm694@cam.ac.uk)